DOCUMENT RESUME

ED 226 065 TM 830 112

TITLE	Testing in the Schools: What Does It Mean?
INSTITUTION	Vermont State Dept. of Education, Montpelier. Div. of Federal Assistance.
SPONS AGENCY	Department of Education, Washington, DC.
PUB DATE	May 81
NOTE	23p.
PUB TYPE	Guides - Non-Classroom Use (055) -- Reports - Descriptive (141)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Achievement Tests; Criterion Referenced Tests; Diagnostic Tests; Elementary Secondary Education; Intelligence Tests; Norm Referenced Tests; *Scores; *Standardized Tests; *Testing; *Test Interpretation; Test Theory
IDENTIFIERS	Elementary Secondary Education Act Title I

ABSTRACT

Because testing, in many different forms, currently plays such an important role in education, Elementary Secondary Education Act, Title I, the Division of Federal Assistance in the Vermont State Department of Education, prepared this brochure to present a general introduction to terms and phrases commonly used in testing and to highlight some of the advantages and disadvantages of intelligence tests, achievement tests, and diagnostic tests. The difference between, as well as the advantages and disadvantages of, norm-referenced and criterion-referenced tests are discussed. The "meaning" of five kinds of test scores are presented: raw scores, grade equivalents, percentiles, stanines, and normal curve equivalents. While this pamphlet attempts to provide an overview on testing, it also points out that testing can be a complicated process that requires a great deal of careful consideration before conclusions can be drawn. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED226065

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

TESTING IN THE SCHOOLS:

What Does It Mean?

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D. Joslyn

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Published By
ESEA, Title I
Division of Federal Assistance
Vermont Department of Education
Montpelier, Vermont 05602

TM 830 12

2

The activity which is the subject of this report was supported in whole or in part by the U.S. Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the U.S. Dept. of Education and no official endorsement by the U.S. Dept. of Education should be inferred.

TESTING IN THE SCHOOLS:
What Does It Mean?

May 1981

Published By
ESEA, Title I
Division of Federal Assistance
Vermont Department of Education
Montpelier, Vermont 05602

STATE OF VERMONT

Governor-Richard A. Snelling

Commissioner of Education
Robert A. Withey

Deputy Commissioner of Education
Edward J. Fabian

STATE BOARD OF EDUCATION

Allen Martin, Chairman	Essex
Alof Carlson	Proctor
Viola Luginbuhl	South Burlington
Louise Swainbank	St. Johnsbury
A. Shirley Tyler	Brattleboro
Thomas P. Whalen	Arlington
Lynn T. Wood	St. Albans

TABLE OF CONTENTS

THE PURPOSE OF THIS PAMPHLET 1

WHY DO WE TEST? 2

WHAT ARE THE MOST COMMON TYPES OF TESTS AND WHAT DO THEY MEASURE? 3
- IQ or Intelligence Tests 3
- Achievement Tests 4
- Diagnostic Tests 5

WHAT IS THE DIFFERENCE BETWEEN NORM-REFERENCED AND CRITERION-REFERENCED TESTS? 7

WHAT DO TEST SCORES MEAN? 11
- Raw Scores 11
- Grade Equivalents 11
- Percentiles 12
- Stanines 12
- Normal Curve Equivalent (NCE) 13

TESTING IS USEFUL, BUT 14

WHAT IS THE VERMONT BASIC COMPETENCY PROGRAM? 16

IN SUMMARY 17

THE PURPOSE OF THIS PAMPHLET

". . . local students tested above the national average in reading"

". . . Thaxter School prepares to administer diagnostic test battery to students"

". . . SAT scores decline for the second straight year"

". . . Ann scored at the 50th percentile in reading and the 95th percentile in mathematics"

". . . Tom is eight months below grade level in reading"

". . . Jon gained 12 NCEs in reading."

We live in a society that is extremely oriented toward testing. Tests are used to report progress; to compare performance; to determine advancement; and to judge success or failure. Testing, in one form or another, appears in almost every phase of life. Testing occurs more frequently, and probably carries more weight, in the field of education than in any other field.

Because testing, in many different forms, currently plays such an important role in education, ESEA Title I, the Division of Federal Assistance in the Vermont State Department of Education, believed it was important to prepare this brochure to present a general introduction to terms and phrases commonly used in testing and to highlight some of the advantages and disadvantages of certain types of tests.

If, after reading this pamphlet, you desire additional information or clarification about testing, please contact either your local school administrator or the Evaluation Consultant, Division of Federal Assistance, Vermont Department of Education, Montpelier, Vermont 05602. Phone: (802) 828-3124

WHY DO WE TEST?

Webster's Dictionary defines the word "test" as: a set of questions or exercises for determining one's knowledge or skills; an examination or trial to determine something's value.

Testing, or determining the level of one's knowledge, has always existed in American education. In the earliest times testing was usually done as part of the daily school routine. Individual students were frequently asked to recite passages or do arithmetic problems on small chalkboards at their desks. The judgment of how well the student had acquired knowledge or mastered the skills being taught was left entirely in the hands of the classroom teacher. Problems with this approach were obvious:

- No way existed to tell whether a teacher was being too harsh or too easy in judging students.
- Tests were not "objective", that is the teacher's judgment played a major role and was subject to biases ranging from student behavior to family background.
- There was no common standard for comparing the performance of one student to other students.

Problems associated with testing and setting objective standards in education became apparent during World War I when it was found that large numbers of young draftees could not read, write, or complete simple arithmetic problems even though they had completed school. It became essential to be able to compare one recruit's skills with another's for placement in training programs. Thus, an effort to develop tests to compare one person's knowledge or skills with those of a group of similar people was undertaken on a large scale. Following World War I, the emphasis on testing continued to grow and testing for the purpose of comparing knowledge and skill levels became a formal part of education and an important tool in improving programs and instruction for students.

Today, schools test to determine how well their students learn what the schools think is being taught. This information is used by teachers, administrators, and educational specialists to help determine how effective educational programs are for students and what additions or changes can be made that will most benefit the student.

Schools use a wide variety of tests in an attempt to determine how well their students are doing. These range from teacher developed classroom tests designed to measure the content of daily lessons to more sophisticated tests developed by large commercial test development companies.

Other than teacher made tests, the most frequently administered tests are intelligence (IQ) tests, achievement tests, and diagnostic tests. The next section of this pamphlet will examine the advantages and disadvantages of each type of test.

WHAT ARE THE MOST COMMON TYPES OF TESTS AND WHAT DO THEY MEASURE?

There are many different types of tests available to measure knowledge and skills. The most frequently used test in the classroom is the one that is developed and administered by the teacher to measure specific information he or she has recently taught. In many respects these are the most important tests a child will take because they allow a teacher to monitor educational progress on a daily basis. However, because teachers must construct and grade so many of these tests, some questions may be poorly phrased and errors may occur in scoring. Furthermore, teacher-made tests do not permit comparisons of a student's performance to other students outside of that particular classroom. For these reasons, commercial publishers hire professional test writers to develop more sophisticated standardized tests and scoring systems.

The most commonly used commercial tests are intelligence tests, achievement tests, and diagnostic tests. Each of these types of tests are used for different reasons. The tests, the reasons for using them, and their advantages and disadvantages will be described below.

IQ or Intelligence Tests

"IQ" stands for intelligence quotient. Intelligence tests are designed to measure a person's "potential" for learning compared to other people his or her own age. Intelligence tests are not meant to measure specific knowledge. Intelligence test scores tend to be fairly constant from year to year but do vary as a function of how motivated a student is to do well on the test, or how well he or she is feeling on a particular day. A child's "potential" for learning may even change as a result of changes in his or her environment or the kinds of educational experiences he or she encounters. IQ tests are not used in ESEA Title I testing.

There is a great deal of controversy surrounding the issue of intelligence testing. As a result, IQ scores are used much less frequently now than they were in the past. Experts can still not agree on a definition of intelligence, whether or not there are different kinds of intelligence, the degree to which intelligence is hereditary, the extent to which tests actually measure intelligence, the amount of cultural bias in the tests, the stability of test scores over time, and the weight which should be given to IQ scores in an overall evaluation of the child. It is important, therefore, that intelligence scores be interpreted by people who are aware of the controversies surrounding these tests and who have other information available about the child.

Remember, intelligence tests, like all other tests, are only one indicator of a student's intellectual development and should never be the only criterion for judging a student's abilities or skills.

Advantages

- IQ scores have been fairly accurate in the past for predicting student success in academic settings.
- Individual IQ tests can be used as one measure for the selection of students for special programs or to identify those in need of special assistance. Normally, an IQ score of 100 is considered "average." The further away from this score (either up or down) a child scores, the greater the possible need for a special program or assistance.

Disadvantages

- Oftentimes intelligence tests do not take into account differences in the cultural or economic backgrounds of students being tested.
- The results of IQ tests are often misinterpreted. IQ scores can and do change. Accurate interpretation can be done only by trained experts.
- Accurate IQ tests can only be given to one student at a time rather than to large groups. This requires much more time for testing and adds to the expense. In addition, most intelligence tests can only be administered by a specialist trained in giving a particular intelligence test.

Intelligence tests are being used less frequently in today's classrooms than in the past. A more commonly used commercial test today is the achievement test.

Achievement Tests

Unlike intelligence tests which were designed to measure a person's "potential for learning", achievement tests are intended to measure a person's general skills in specific academic areas such as vocabulary, reading comprehension, arithmetic computation, spelling, social studies, or science. These tests are often referred to as "survey" tests because they do not try to determine everything a student may know about a subject. Usually, achievement tests are constructed by test publishers using experts from universities, textbook writers, and curriculum specialists. These people examine what is being taught at different grade levels across the country in areas such as reading and mathematics. Based on this information, these experts develop test questions to measure generally how well students are doing in each academic area.

Achievement tests are commonly used in selecting students to par-

ticipate in ESEA Title I. This is because student Title I participation is contingent upon a student's not achieving in basic skills on a par with his/her peers. These tests are also commonly used for reporting the gains students show from ESEA Title I assistance.

While achievement tests are popular among educators they, like all tests, should be looked upon as only one indicator of how well a student or educational program is succeeding.

Because achievement tests are "general" measures of an academic area, not all the questions on the test will precisely match what is being taught in your child's school or classroom. These tests are useful for getting a *general* picture of how well a group of students or an educational program is functioning in a specific subject matter area.

Advantages

- Achievement tests allow parents and educators to compare how well their students and schools are performing in terms of general knowledge in such areas as reading, mathematics, science, etc., with other students and schools from across the country.
- Achievement tests are usually administered to large groups of students and, therefore, are relatively inexpensive and less time consuming than intelligence tests.
- Achievement tests can be easily administered and interpreted by classroom teachers.

Disadvantages

- The questions on achievement tests do not measure precisely what is being taught in a classroom or school. There may be questions on the test that measure information that is not being taught at that grade level or in that school. There also may be information or skills being taught in a grade or school for which there are no questions on the achievement test.
- Only content areas which can be measured by multiple-choice test items are included in the test.
- Achievement tests cannot be used to pinpoint the strengths and weaknesses of individual students in various subject matter or skill areas.

If achievement tests provide *general* information on how well students or programs are performing, what test can be used to pinpoint problem areas? The answer is a diagnostic test.

Diagnostic Tests

While the achievement test is intended to measure general knowledge or skill in a subject area, the diagnostic test is intended to identify

specific problem areas within that subject matter. For example, an achievement test in arithmetic computation may simply show that the student is performing poorly in that general area. A diagnostic arithmetic test might show that the student can do everything but carry and borrow in addition and subtraction problems.

For the most part, diagnostic tests are the most popular test tool used by teachers and educational specialists. These tests allow teachers to identify the specific strengths and weaknesses a student has and then plan a precise program to overcome those weaknesses. Diagnostic tests are commonly used for ESEA Title I testing to assist teachers both in diagnosing problems and in checking for student progress.

Diagnostic tests, unlike achievement tests, are usually designed to measure only reading or mathematics and not subject areas such as social studies and science.

Advantages

- Unlike achievement tests, diagnostic tests provide the teacher with a detailed picture of a student's strengths and weaknesses.
- Results allow the teacher to develop precise instructional plans to compensate for weaknesses and to build on strengths.
- Usually diagnostic tests can be administered to large groups of students which make them less expensive and time consuming than intelligence tests.

Disadvantages

- Diagnostic tests usually provide information on only one subject area at a time, which is usually either reading or mathematics.
- Diagnostic tests require more training to administer and interpret than achievement tests but less training than intelligence tests.
- Diagnostic test information is of little value unless someone can take that information and develop a specific educational program to overcome student weaknesses.

In summary, let us review what we have learned about tests that are commonly used in education:

- Teacher made tests, although extremely valuable in measuring student progress, are generally less sophisticated and do not permit comparisons across grades or schools.
- Intelligence tests are being used less frequently in schools, usually require an expert to administer and interpret, are given to one student at a time, and measure "potential for learning" rather than specific knowledge in a particular content area.

- Achievement tests are the most commonly used commercial test. they measure "general" knowledge levels in several subject matter areas. they give an overall picture of how well a student or program is doing. Not all of the questions on an achievement test match what is being taught in a classroom or school.
- Diagnostic tests are probably the most popular tests among teachers. they identify specific stuoent strengths and weaknesses, usually in math or reading. the information can be used to develop precise educational plans for students to overcome weaknesses.

Finally, tests are only one indication of how well a student or educational program is functioning and this information, by itself, should not be the sole criterion for judging performance.

For example:

Judy W. has a record of scoring poorly on any type of commercially developed test. In talking with her we find out that the "pressure" created by the way these tests are given frightens her and that she finds she can not concentrate on the questions. How else can we measure how well Judy is doing in school? We could look at other test scores to see if she has always had this problem in taking tests or whether it is a recent development. we could look at her grades in various subjects. talking with her teachers would give an indication of her ability. we could examine samples of her daily work, or we could observe her performance in the classroom periodically. While none of these alternatives to testing give us a broad basis for comparing Judy to other students, they do allow us to make general statements about her skills and ability.

WHAT IS THE DIFFERENCE BETWEEN NORM-REFERENCED AND CRITERION-REFERENCED TESTS?

We have talked about the different types of tests usually found in schools. Since achievement tests are the most commonly used tests in education, and most schools use achievement tests at one time or another, it is important to understand at least two different ways these tests are used to make comparisons of students.

We can see how well a student did on an achievement test compared to similar students from across the country. In test jargon, a test that uses this kind of comparison (an individual's performance compared to a group's performance) is called a "norm-referenced test".

We can also compare how well a student did on an achievement test

by matching his or her performance to predetermined criteria. For example, we might say that we expect all fourth grade students to be able to pass 9 out of 10 addition problems on a test. If the student cannot do this, he or she will be given special assistance. In this case, the student's performance on the test is being compared to a predetermined expectation or criterion, and, therefore, we call this kind of achievement test a "criterion-referenced test."

Let us examine these two kinds of comparisons from achievement tests more closely.

The "group comparison or norm-referenced test" is simply a test whose questions have been given previously to large numbers of students from all over the country. When test publishers develop an achievement test, they ask schools across the nation to give the test to their students. In return, the schools are not charged for the test.

By giving the test to large numbers of students from urban, suburban, and rural schools in regions throughout the country, the test publisher hopes to get a cross-section of student scores that reflect how students of different skill levels perform on the test. Based on these scores, the test publisher then has an idea of how an "average", "above average", or "below average" student will score on this test. It is this average or "normal" population's performance on the test against which your school's students are compared.

The actual comparison is made by taking the number of correct answers your student received and going to a table in the test booklet that provides scores of how well the "normal population" did on the same test.

The reason the test publisher gives the test to so many students initially is to get a better picture of how well the "typical or average" student can be expected to score and to insure that students from varied backgrounds have been included in the "normal population". For example, it would not be fair to test students from poor, rural areas on an achievement test which had been tried out on students from rich, suburban areas. We can assume that wealthier suburban schools and communities have more resources available both in and out of school that affect how well their students do on tests. Rural students often do not have these same resources available. Therefore, it is not fair to compare how well they perform on tests "normed" on just students from wealthy suburban areas.

On a "norm-referenced test" then, comparisons can be made between your child's performance and that of other students of similar age, grade level, or background from across the country. Scores on a "norm-referenced test" may also be used to determine how well a school's or grade's performance compares with similar schools or grades across the country.

Advantages

- Norm-referenced tests provide a means of compar-

ing the performance of individual students or educational programs to other similar students or programs from across the country.
- Results from the publisher's "norming group" can be used as a benchmark for determining how much learning has taken place and for identifying areas of general weakness in curriculum.

Disadvantages

- Norm-referenced tests always judge a student's performance relative to the performance of the students in the "normal population." This norm population may not be made up of similar students and so comparisons would not be fair.
- The results of "norm-referenced tests" are often over-interpreted and given more weight than other indicators of learning. While it is fair to use the norm-referenced test results for general comparisons, you need to remember that it does not perfectly match what is being taught in your local schools.
- Data collected from the "norming group" quickly becomes out-of-date because of changes in curriculum as well as in society as a whole. For example, it probably is misleading to compare the performance in reading of fourth graders in 1979 to how fourth graders in the "norm population" did in 1969. Teaching methods changed; textbooks were improved; students were exposed to vastly different experiences; social or parental pressures may have changed — all these factors probably contributed to making the 1969 norms outdated.

An alternative to comparing a student against how another group of similar students did on a test is to compare his/her performance against a predetermined criterion or standard. This type of test is called a "criterion-referenced test."

A criterion-referenced test compares the student to a set of criteria rather than to the performance of other students. Questions are usually arranged to measure skills in some type of sequence from simplest skills to the more complicated. If a student can pass all or most of the items we say he or she has "mastered" the skills. If the student begins to fail questions, then we say that mastery of that skill has not been attained, and that is where instruction begins. On a criterion-referenced test there is usually no comparison with how other students did on the same questions. We are interested in only what each individual student knows.

On a math criterion-referenced achievement test, we may expect a fourth grader to be able to do 4 of the first 6 items correctly and a fifth

grader to do 10 of the first 14 items. If a student in either grade fails to obtain the necessary number of correct items ("the criterion") then a special program of instruction can be provided for them. This allows students to learn at their own pace and to be operating at different levels in different subject areas without specific references to grade levels.

Advantages

- Student achievement is judged on how well that student performed a desired skill rather than by comparing his or her performance to another group of students functioning at a different grade level.
- Teachers can obtain information about how well individual students have mastered specific skills and use that information to develop individual programs of study for those students.

Disadvantages

- In most cases, there is no way of knowing how a student's score on a criterion-referenced test compares with a national average.
- Test results usually cannot be summarized in a simple score.

In summary, there are two basic kinds of achievement tests: norm-referenced tests and criterion-referenced tests. The difference between these tests is how they evaluate a student's performance.

A norm-referenced test compares a student's test performance to how well a "group or normal population" did on the same test. This type of test allows a school to see how well its students or programs are doing compared with similar students or programs across the country. If the "norm-population" is not similar to your students then the comparisons will be misleading. The results of ESEA Title I assistance are usually reported on the basis of norm-referenced tests.

A criterion-referenced test compares the student to a set of criteria or expectations in terms of skills to be mastered. In many ways criterion-referenced tests, like diagnostic tests, allow the teacher to see which skills have been mastered and which have not and to plan specific educational programs for the student. Many ESEA Title I teachers use criterion-referenced tests to monitor student progress. The results of some criterion-referenced tests such as the PRI-DMI can be converted, by the test publisher, to a report on how well the child did in comparison to other students across the country.

WHAT DO TEST SCORES MEAN?

We have discussed the types of tests used in schools and looked specifically at achievement tests, but tests are of no value unless we can understand what the scores produced really mean. We will look at five kinds of test scores: raw scores, grade equivalents, percentiles, stanines, and NCEs. These scores are usually found only on norm-referenced tests since they represent different ways of comparing one student's performance with that of a group of peers.

Raw Scores

A raw score is simply the number of questions the student answered correctly. Because the number of questions vary between tests, the raw score itself does not have any value in making comparisons. For example, if a fifth grade student gets 21 out of 50 questions correct on a reading test and then gets 12 questions out of 25 correct on a math test, what do we know about the student's performance in reading and math? The answer is nothing, since we don't know how hard the questions are for a fifth grader. We use the raw score to go to the test publisher's tables and convert to scores that let us compare the child's performance to that of other fifth graders. One of those scores is the grade equivalent.

Grade Equivalents

Grade equivalent scores are based on a division of the school year into nine months.

Beginning									End of
Sept.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June
0	.1	.2	.3	.4	.5	.6	.7	.8	.9

If a student is functioning at grade level, that student is demonstrating the same level of skill or knowledge as other students at that grade level. In other words, he or she is getting the same number of items correct on a test as the average student at the same grade level. For example, a third grade student who is tested in reading in October and who is said to be scoring at the 3.1 grade level (third grade, first month) is at grade level. A student scoring 3.0 (third grade, no months) is slightly below grade level; 3.2 (third grade, second month) is slightly above grade level. If a third grade student scores 10.3 (tenth grade, third month) on a reading achievement test, does this mean he or she should be in the tenth grade? No, it simply means that the student is achieving and probably could do more challenging work, though not necessarily at the tenth grade level. The third grade reading test, in this case, was never given to tenth graders. The test publisher arrived at the tenth grade score through a statistical formula rather than actually giving the test to tenth graders.

The major disadvantages to the grade equivalent score are that it cannot be added and averaged accurately and that it assumes that the same amount of learning occurs in each month of the school year. We know that is not true. Therefore, changes in grade equivalent scores are not necessarily accurate indicators of student progress.

If grade equivalents are not good indicators of progress what about percentiles?

Percentiles

A percentile is another type of score that is derived from the raw score (number of correct responses). The percentile is a scale from 1 to 99, with the 50th percentile being considered average.

If a student receives a raw score of 20 on his/her reading achievement test and we look this up in the test publisher's table, we find it converts to a percentile score of 60. What this means is that approximately 40% of the students in the "norm population" scored higher than this student in reading and approximately 60% of the students in the norm population scored lower in reading.

Obviously, the higher above the 50th percentile students score, the better they are doing, and the further below the 50th percentile they score, the greater the need is for special assistance.

Although there are problems with adding and averaging percentiles, these problems are not as great as they are with grade equivalents. Percentiles provide a much more accurate description of student progress than grade equivalents.

Another commonly used test score is called the stanine.

Stanines

A stanine is another indicator of a student's rank relative to the norm population and is also derived from the test publisher's table by using the raw score.

Stanines are a scale from 1 to 9. Stanine scores of 1, 2, or 3 are considered to be below average scores; scores of 4, 5, or 6 are average; and stanine scores of 7, 8, or 9 are above average scores.

While the stanine, unlike the grade equivalent and percentile scores, can be added and averaged accurately, its disadvantage is that it is not a very precise indicator of student progress. For example, a student may get 27 questions correct the first time he/she takes the test. This number of correct questions may just barely be enough to get a stanine score of 3. At the end of the year the student takes the same test and this time gets 32 items correct. This number of correct questions is only one away from a stanine score of 4 *but* it is still one short, so the student still has a stanine score of 3. The student had a stanine score of 3 when he or she started and still has a stanine score of 3 at the end of the year. Does this mean no learning took place? Obviously not, because we saw from the example that the student answered more questions cor-

rectly the second time than he or she did the first time. What it means is simply that the stanine score is not exact enough to accurately reflect student progress.

NCE

A score which is technically very similar to a stanine, but looks like a percentile, is the NCE, or normal curve equivalent. While stanines range from 1 to 9, NCEs range from 1 to 99. If one translated stanine scores into NCEs, scores of 1 to 35 NCEs would be considered below average; scores of 34 to 66 would be average; and scores of 67 to 99 would be above average. Because there are 99 points on the scale instead of 9, NCEs provide a much more precise measure of where a student is, and how much he or she progresses relative to other students. Like stanines, NCEs can be averaged to provide an accurate picture of group performance. In fact it was for the purpose of summarizing nationwide ESEA Title I achievement data that NCEs were initially developed. NCEs are not often used to report test results of individual students, but are used to report the ESEA Title I results that a school district has achieved.

The following excerpts on NCE's are taken from Technical Paper No. 2 by G. Kasten Tallmadge entitled *Interpreting NCE's — ESEA Title I Evaluating and Reporting System* published in October 1976 by the Office of Education.

"*NCEs are like percentiles.* Both an NCE of 50 and a percentile of 50 are exactly average. While NCEs do not match percentiles at other points (except for 1 and 99), the analogy is quite useful when trying to describe achievement gains measured in NCEs. While it is not strictly correct to talk about NCE gains as if they were percentile gains, it will probably facilitate communication and enhance understanding to do so. This is particularly true since most people tend to think of percentiles as if they were an equal-interval scale and would be somewhat confused to learn that a gain from percentile 5 to percentile 10 is almost exactly twice as big as a gain from percentile 15 to percentile 20.

"*An NCE of 50 is at grade level.* Regardless of the time of year at which testing is done and the grade level tested, a properly derived NCE score of 50 will always be the national average for that grade level and month. Being average means being exactly at grade level. NCEs below 50 signal below-average achievement levels or below-grade-level performance. An NCE of 30 is exactly the same distance below grade level at every grade while being "a year below grade level" has a different meaning at each grade. Finally, an NCE of 30 is always exactly twice as far below grade level as an NCE of 40 while being "two years below grade level" is *never* twice as much as being one year below grade level (believe it or not)!

"*An NCE gain of zero means that the Title I project produced no gain.* A zero NCE gain does *not* mean that the student or group of students

learned nothing between pretest and posttest. They almost certainly answered more items correctly at the end of the instructional period than at the beginning. The zero NCE gain simply means that the amount of learning was precisely what would have been expected had there been no Title I project — in other words it means that the Title I project added exactly nothing to the regular school program.

"*All NCE gains greater than zero are good!*" Whenever the evaluation shows an NCE gain greater than zero, it means that the Title I pupils profitted from participating in the project. In general, the larger the NCE gain, the more effective the project. It is not possible, however, to designate any specific NCE gain as the criterion for exemplary or outstanding projects. A cost-effectiveness criterion seems more appropriate. Assuming that the same number of dollars were spent, for example, a 4-NCE gain produced in a treatment group of 200 pupils might be considered as good as an 8-NCE gain produced in a treatment group of 100 pupils."

In summary, we now know that raw scores, or the number of items a student gets correct on the test have little value for comparison purposes but are the key to getting other test scores from the publisher's tables. Grade equivalents are not accurate for describing student progress because they cannot be added or averaged accurately and they erroneously assume that equal learning takes place in each month of school.

Stanines can be averaged accurately but are not precise enough indicators of student progress. Generally, the best test scores for showing student growth or progress is the percentile or the NCE.

Finally, we should emphasize that caution should be used in interpreting test scores. If you, as a parent, are unsure about the meaning of your child's test scores, seek advice from your school. *Remember, test scores are only one indication of a student's performance.* We should not discount other factors that serve as evidence of student performance such as subject matter grades, examples of the student's daily work, teacher appraisal, or observation of the student performing in class.

TESTING IS USEFUL, BUT:

Testing is a very valuable tool for education, parents, and the community when it comes to making decisions about programs for students and schools. When properly used, test information can help teachers create educational plans designed to overcome areas of academic weakness or to build on specific strengths. Test information can be used by school administrators and citizens to make decisions about where resources should be applied to meet educational needs.

Obviously, test information can also be abused and to prevent this we offer the following cautions:

- Testing is only *one* indication of how well a student or school program is performing. Other information should be included along with test information before judgments are made about students and programs.

  For example, information about the attitudes and behavior of the students in school is probably as important as how well they do on achievement tests. Information about the resources available to provide the instruction is also important.

- If achievement tests do not match the material taught to the student, then drawing conclusions from such tests can be misleading, and the results will be of little or no value.

  For example, if your school uses a traditional mathematics curriculum and the new achievement test you have selected tests modern math concepts, the test will not accurately represent how your students are doing in math. It is extremely important that the content of the achievement test matches as closely as possible what is being taught.

- Some students do not perform well on tests; for them, a test is not an accurate reflection of achievement.

  For example, some students become so apprehensive when it comes time to take a test that they "freeze" or "go blank". There is no doubt that testing creates pressure on students. For students who cannot take tests, we must use other indicators such as teacher grades, observation of classroom performance, and examples of work. Efforts should also be made to reduce test anxiety for students by telling them the purpose and use of the tests they are taking and providing them with more experience in taking tests.

- Do not "over-interpret" test results. Be aware of the limitations of the test you are using and the test scores that are being reported.

  For example, our fourth grade scored this year at the 48th percentile in reading but every other year they have been at the 52nd percentile. Something must be wrong! To begin with, the 48th percentile is not significantly below average in terms of performance and, for that matter, the 52nd percentile is not that high above average. Do not generalize about the quality of an entire school based on the performance of how one grade scored on one achievement test.

- A test must be properly administered in order to obtain accurate results.

  For example, giving an achievement test just after students have returned from vacation will probably not give you the best results. Also, testing students for long periods of time during the day can lead to poorer scores because students get tired.

  Teachers should be familiar with directions for giving the test. If a test is to be given to students for 30 minutes, the

teacher should not extend the time to 45 minutes because she knows if the students had more time they would do better. The norm population that originally took the test had only 30 minutes. If valid comparisons are to be made, that is all the time that can be allowed. Following directions on a test is essential.

Teachers should not teach the specific questions on a test to students. Naturally, they may teach the subject matter or content that those questions are intended to measure.

Since the State of Vermont has introduced a basic competency program that allows school districts the option of determining what methods will be used to assess student skill levels (locally developed tests, norm-referenced tests, criterion-referenced tests, etc.), it seems important to provide a brief explanation of that program.

WHAT IS THE VERMONT BASIC COMPETENCY PROGRAM?

All of the schools in the State of Vermont are probably doing some type of testing similar to what has been described in this pamphlet. In addition, all are involved in the Vermont Basic Competency Program. The purpose of this program is to insure that students graduating from Vermont schools have obtained minimal or basic mastery of skills in the areas of reading, writing, listening, speaking, computation, and reasoning.

The Basic Competency Program allows the use of a flexible testing system. The State of Vermont has already stated what the basic skills should be in each area, but it leaves the measurement of those skills up to each community. Some communities wish to assess these skills via norm-referenced or criterion-referenced tests available from commercial publishers. Most communities, however, have chosen to develop their own test items. While the Vermont Basic Competency program establishes the minimum acceptable level of skill for students and schools in reading, language arts, arithmetic, and reasoning that must be demonstrated before a diploma or advancement is awarded, the methods for measuring these skill areas are left up to the schools. Schools may elect to use a commercially developed test or to develop tests of their own or to combine both approaches.

While the testing process used by each school district participating in the Vermont Basic Competency Program is different and therefore, does not permit comparisons to be made among school districts, it does provide the district with specific information about how well its students and schools are performing in the basic skill areas identified by the state.

IN SUMMARY

We have examined the kinds of tests that are most frequently used in schools and what various types of test scores mean. What we can conclude is that testing has been, and probably will continue to be, a very integral part of the educational process. It is, therefore, important for parents and teachers to become more familiar with testing and the issues associated with the use of tests.

While this pamphlet has attempted to provide an overview on testing, it has also pointed out that testing can be a complicated process that requires a great deal of careful consideration before conclusions can be drawn.

Tests are a tool that can help parents and teachers improve educational programs for students, but they are only *one* tool. Unless they are used wisely, they can easily distort the overall picture of a child's educational development.